DOCUMENT RESUME

ED 177 167 TM 008 516

AUTHOR Smith, Sandra E.; And Others
TITLE When Is An F-ratio Not an F?
PUB DATE Mar 78
NOTE 10p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978)
AVAILABLE FROM Department of Educational Psychology, University of Wisconsin, 1025 West Johnson Street, Madison, Wisconsin 53706

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Analysis of Variance; *Hypothesis Testing; *Reliability; Research Problems; Statistical Analysis; *Tests of Significance
IDENTIFIERS *F Test

ABSTRACT

A correction of the standard F-ratio for unreliability of the dependent measure has recently been proposed by Winne; the rationale is analogous to that of correcting a correlation for attenuation. However, there are two problems associated with Winne's correction of which potential users should be aware. First, the corrected statistic, F*, has been incorrectly derived; and second, it has been assumed without evidence that F* is distributed as is the common F-ratio. These problems are illustrated (via Monte Carlo procedures) and solutions are suggested. (Author/JKS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

PAPER SESSION
ID # D-27

ED177167

When is an F-ratio not an F ?

Sandra E. Smith, Michael J. Subkoviak, Joel R. Levin

University of Wisconsin

Presented at the annual meeting of the American Educational Research Association,
Toronto, 1978
Printed in U.S.A.

TM008 516

Abstract

A correction of the standard F-ratio for unreliability of the dependent measure has recently been proposed by Winne (1977) — the rationale being analogous to that of correcting a correlation for attenuation. However, there are two problems associated with Winne's correction of which potential users should be aware. First, the corrected statistic, F*, has been incorrectly derived; and second, it has been assumed without evidence that F* is distributed as the common F-ratio. The present paper seeks to illustrate these problems (via Monte Carlo procedures) and to suggest solutions.

When is an F-ratio Not an F ?

A correction of the F-ratio for unreliability of the dependent measure has been recently proposed by Winne (1977). The rationale is similar to that for correcting a correlation for attenuation. Since inferences drawn from statistical tests are influenced by unreliability of the dependent variable, if appropriate allowances are made for this unreliability, an increase in the precision and power of assessment of treatment effects should occur (see, for example, Cleary, Linn, & Walster, 1970; Levin & Subkoviak, 1977). For these reasons, such a correction might be useful for certain purposes. Our concern is that some researchers currently employing Winne's correction may not be aware of two problems associated with it. First, the corrected statistic, F^*, has been incorrectly derived. In addition, it has been assumed without theoretical or empirical justification that F^* is distributed as the common F statistic — something not likely to be the case given the ingredients of F^*.

The F^* statistic derived by Winne is the common F statistic divided by a sample estimate of the reliability coefficient, i.e., $F^* = F / \hat{\rho}_{xx'}$. However, the correct derivation leads to the following result:

$$F^{**} = \left[\frac{F - (1 - \hat{\rho}_{xx'})}{\hat{\rho}_{xx'}} \right].$$

The presence of the random variable $\hat{\rho}_{xx'}$ suggests that neither F^* nor F^{**} would be distributed as the common F. Thus, the significance of F^* or F^{**} should not be determined by using the common F-tables as Winne suggests. If, in fact, these statistics are not distributed as F, then new tables would have to be derived.

Method

A Monte Carlo study was conducted to illustrate the extent to which F* and F** deviate from the distribution of the common F-statistic, when the null hypothesis is true (i.e., when there are no treatment differences).

Observed and true score pairs, (X,T), were sampled from a bivariate normal population. Since the expected values of true and observed test scores are equal, μ_T and μ_X were set equal to one another. In particular, $\mu_T = \mu_X = 50$ was used in all cases. σ_X^2 was set equal to 100, and σ_T^2 varied according to the desired $\rho_{XX'}$. For example, if $\rho_{XX'} = .95$, then $\sigma_T^2 = \sigma_X^2 \rho_{XX'} = (100)(.95) = 95$.. Notice also if $\rho_{XX'} = .95$, then the correlation between X and T is $\rho_{XT} = \sqrt{\rho_{XX'}} = \sqrt{.95} = .97$.

The following quantities were varied:

1. $\rho_{XX'} = \rho_{XT}^2 = .55, .75, .95$ (the reliability coefficient).
2. N = 10, 20, 30 (the size of the sample drawn from the bivariate normal distribution of X and T).
3. k = 2, 4 (the number of treatment groups).

One thousand samples of size N were drawn from k bivariate normal distributions of X and T , and the F, F*, and F** statistics were computed for each sample of X scores.

In each case the pooled sample correlation between X and T, $\hat{\rho}_{XT(pooled)} = \left[\frac{\widehat{COV}(X,T)_{(ave.)}}{\sqrt{\widehat{VAR}(X)_{(ave.)} \widehat{VAR}(T)_{(ave.)}}} \right]$, was computed and then squared for use in the F* and F** statistics. The simulated F* and F** distributions, as well as the number of false rejections obtained at standard significance levels (α = .05, .025, .01), were compared with those associated with the

usual F distribution. The distributional properties of F, F*, and F** were also compared. Specifically, the first few moments of each distribution were computed and compared, i.e., means, variances, skewness, and kurtosis.

Results

The distributional forms of F* and F** were found to differ considerably from F, especially for small values of $\rho_{xx'}$. For example, when k=2, N=10, and $\rho_{xx'}$=.55 the following percentages of rejections of the null hypothesis were obtained for the nominal α-level = .025 :

	Nominal α	Actual α				
		Simulated F	F*	F**	$F^*(\rho_{xx'}\text{ fixed})$	$F^{**}(\rho_{xx'}\text{ fixed})$
CASE 1.	.025	.027	.114	.084	.083	.07

But when $\rho_{xx'}$= .95 :

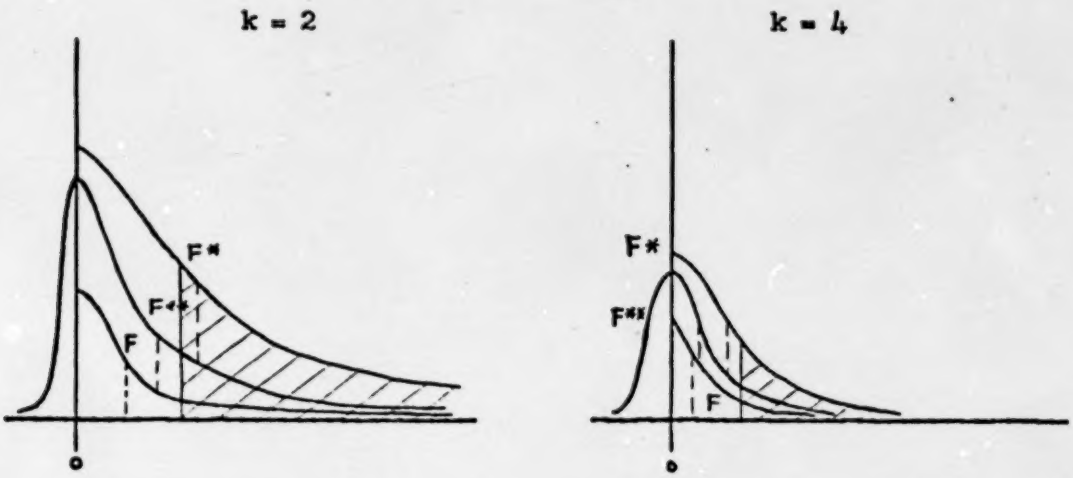
	Nominal α	Actual α				
		Simulated F	F*	F**	$F^*(\rho_{xx'}\text{ fixed})$	$F^{**}(\rho_{xx'}\text{ fixed})$
CASE 2.	.025	.025	.029	.027	.029	.028

If an experimenter were to compute an F* statistic in CASE 1. and unknowingly compare his or her result to the standard F-tables, he or she would have a nominal Type I-error of 2.5%, when the actual Type I-error can be shown to be 11.4 %. Had the experimenter computed F** and made the same mistake he or she would have a Type I error of 2.5 % (nominally) but 8.4 % (actually).

In CASE 2. comparing a computed F* to the standard F-tables yields a nominal Type I-error of 2.5 % when the actual Type I-error is 2.9 %. Comparing a computed F** to the standard F-tables yields a nominal Type I-error of 2.5 % when the actual Type I-error is 2.7 %. The percentages of rejections of the null hypothesis are also given for F* and F** when $\rho_{xx'}$ is fixed at the .55 or .95 level.[1] In general, it was found that for given levels of $\begin{cases} k, \\ N, \\ \rho_{xx'} \end{cases}$:

$$\left[\text{Type } \underline{F} \text{ I-error}\right] \lessapprox \left[\text{Type } \underline{F^{**}} \text{ I-error}\right] \lessapprox \left[\text{Type } \underline{F^{*}} \text{ I-error}\right].$$

In comparing the moments of the distributions, it was found in general that the estimated expected values of $\underline{F}$, $\underline{F}^*$, and $\underline{F}^{**}$ were related as follows : $E(\underline{F}) \leq E(\underline{F}^{**}) \leq E(\underline{F}^*)$. A general picture of the obtained respective distributions would look something like :



As one would expect, as N increased or as k increased the differences among these means decreased. For example,

CASE	Theoretical E(F) $\nu_2/(\nu_2-2)$	Obtained Values E($\underline{F}$)	E($\underline{F}^{**}$)	E($\underline{F}^*$)
k = 2, N = 10, $\rho_{xx'}$ = .95	18/16 = 1.125	1.126	1.153	1.206
k = 2, N = 30, $\rho_{xx'}$ = .95	58/56 = 1.036	1.022	1.028	1.079
k = 4, N = 10, $\rho_{xx'}$ = .95	36/34 = 1.059	1.045	1.050	1.104

The variances followed the same pattern as the expected values. The $\underline{F}^*$ and $\underline{F}^{**}$ distributions were in general more positively skewed than the $\underline{F}$ distribution, with $\underline{F}^*$ being consistently the most positively skewed. As N or k increased, the skewness decreased. Concerning the kurtosis, the $\underline{F}^*$ distribution was found to be the most peaked followed closely by the $\underline{F}^{**}$ distribution, then lagging further behind — the $\underline{F}$ distribution. As N or k increased the peakedness decreased.

Conclusions

It is evident that the proportion of false rejections obtained at standard significance levels differs from distribution to distribution.

Hence, if F^* or F^{**} are used in conjunction with standard F-tables, the nominal and actual Type I-error probabilities will not coincide. This says nothing about the likely discrepancy between nominal and actual Type II-errors. For significance testing with F^{**} (the correctly derived statistic), entire "F^{**}-tables" would have to be computed. (We have essentially computed a few of these entries by our simulations.) To calculate power and Type II-error probabilities would require knowledge of a "non-central F^{**} distribution".

One wonders whether Winne's logic may be faulty. His derivation involves subtracting measurement error variance, $\hat{\sigma}_e^2$, from both the numerator and denominator of the conventional $F = MS_B/MS_W$:

$$F^{**} = \frac{F\,(MS_W) - \hat{\sigma}_e^2}{MS_W - \hat{\sigma}_e^2} \quad .$$

This may not be the proper way to estimate what the ANOVA outcome would be if the dependent variable were measured without error.

A possible alternative open to the researcher who has reliability information at hand might be to use the reliability coefficient to estimate true scores from observed scores (see, Lord & Novick, 1968, pp. 64-68). Specifically, the true scores might be estimated as follows:

$$\hat{T}_i^{(k)} = \hat{\rho}_{xx'} X_i^{(k)} + (1 - \hat{\rho}_{xx'})\,\bar{X}^{(k)},$$

where i denotes an individual in group k; $\hat{T}$ is the estimated true score; X is the observed score; $\hat{\rho}_{xx'}$ is a common reliability estimate (for example, see Levin & Subkoviak, 1977); and $\bar{X}$ is the mean of the group of X scores. Subsequent analysis might then be carried out utilizing the estimated true scores instead of the observed scores. Whether or not standard ANOVA would be appropriate on the estimated true scores is a question for further investigation. So, for that matter, is the philosophical issue of whether this or any other unreliability-adjusting procedure is logically defensible (including the kind of abuse such procedures are likely to encounter in the hands of educational researchers).

References

Cleary, T. A., Linn, R. L., & Walster, G. W. Effect of reliability and validity on power of statistical tests. In E. F. Borgatta and G. W. Bohrnstedt (Eds.), Sociological Methodology. San Francisco: Jossey-Bass, 1970.

Levin, J. R., & Subkoviak, M. J. Planning an experiment in the company of measurement error. Applied Psychological Measurement, 1977, 1, 331-338.

Lord, F. M., & Novick, M. R. Statistical Theories of Mental Test Scores. Addison-Wesley Publishing Company, Inc. , 1968.

Winne, P. H. A correction for unreliability of the dependent variable in the analysis of variance. A paper presented at the annual meeting of the American Educational Research Association-New York, 1977.

Footnote

[1] In these F-ratios, the known population values of $\rho_{xx'}$ replaced the data-derived $\hat{\rho}_{xx'}$ estimates. In general, using population rather than sample reliabilities did not appear to make much difference, however.